Данный раздел посвящён сравнению моделей LDA и ARTM. В экспериментах, сравнивающих BigARTM с Gensim и VW.LDA, мы показали, что алгоритм Online PLSA со сглаживающим регуляризатором и Online VB LDA работают схожим образом. Поэтому в этом эксперименте будут оцениваться характеристики PLSA со сглаживающим регуляризатором (который мы далее будем называть LDA) и ARTM (суть PLSA с набором регуляризаторов).

Текстовая коллекция Все наши эксперименты проводились на корпусе английской Википедии [1] , объём которой $|D| \approx 3.7 \times 10^6$ документов. Словарь имеет размер $|W| \approx 10^5$, общая длина коллекции в словах $n \approx 577 \times 10^6$.

Параметры эксперимента В этом эксперименте мы будем пользоваться следующими функционалами качества моделирования:

- Перплексия на контрольной выборке [2].
- Разреженность матрицы Φ.
- Разреженность матрицы Θ документов обучающей выборки.
- Характеристики ядер тем (размер, чистота, контрастность) ([**?**] ССЫЛКА НА НУЖНУЮ ПУБЛИКАЦИЮ КВ!!!).

Обе модели будут иметь следующий общий набор параметров, с которыми будет запускаться BigARTM: 1 проход по коллекции [3] , 10 проходов по каждому документу, 100 выделяемых тем. Матрица Θ, построенная на предыдущем проходе по документу, используется в качестве начального приближения на текущем. Параметры обновления матрицы Φ, κ и τ_0, равны 0.5 и 64 соответственно [4] . Порог $p(t|w)$ для ядровых функционалов — 0.25,. Размер батча равен 10000, обновления модели производится каждые батч.

Параметры LDA $\alpha = \beta = \dfrac{1}{|T|}$.

Регуляризатор для ARTM, представляющий собой смесь разреживания и декорреляции тем, описывается формулой

$$R(\Phi, \Theta) = -\beta \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} - \gamma \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max. \quad (1)$$

[1] Коллекция была получена с помощью gensim.make_wikicorpus.

[2] Объём контрольной выборки, на которой перплексия измерялась в ходе прохода по коллекции — 10 тыс. документов. Кроме того, была измерена результирующая перплексия на выборке из 100 тыс. документов.

[3] Подразумевается один полный проход по всей коллекции и повторный проход по первым 1.5×10^5 документам для уточнения их распределений.

[4] Как это было в экспериментах в **??** РАЗДЕЛ ПРО СРАВНЕНИЕ БИБЛИОТЕК!!!

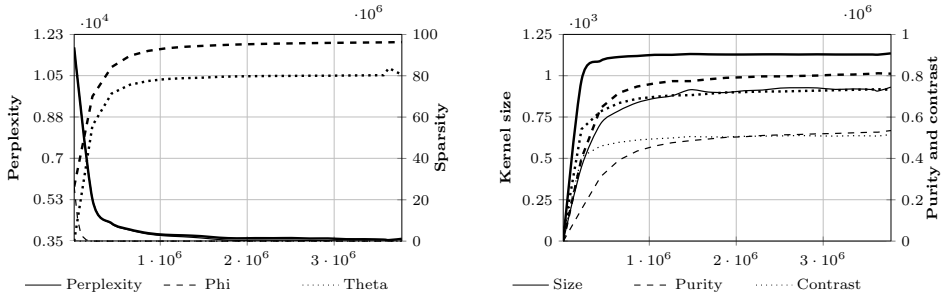

Рис. 1: Comparison of LDA (thin) and ARTM (bold) models. X axis is a number of processed documents.

Таблица 1: Comparison of LDA and ARTM models. Quality functionals: $\mathcal{P}_{10k}$ $\mathcal{P}_{100k}$ — hold-out perplexity on 10.000 and 100.000 documents sets, $\mathcal{S}_\Phi$, $\mathcal{S}_\Theta$ — sparsity of Φ and Θ matrices (in %), $\mathcal{K}_s$, $\mathcal{K}_p$, $\mathcal{K}_c$ — average topic kernel size, purity and contrast respectively.

Model/Functional	$\mathcal{P}_{10k}$	$\mathcal{P}_{100k}$	$\mathcal{S}_\Phi$	$\mathcal{S}_\Theta$	$\mathcal{K}_s$	$\mathcal{K}_p$	$\mathcal{K}_c$
LDA	3499	3827	0.0	0.0	931	0.535	0.516
ARTM	3592	3944	96.3	80.5	1135	0.810	0.732

Отсюда получаются формулы М-шага

$$\phi_{wt} \propto \Big(n_{wt} - \beta \underbrace{\beta_w[t \in T]}_{\substack{\text{sparsing}\\ \text{topic}}} - \gamma \underbrace{[t \in T]\, \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws}}_{\text{decorrelation}}\Big)_+; \tag{2}$$

$$\theta_{td} \propto \Big(n_{td} - \alpha \underbrace{\alpha_t[t \in T]}_{\substack{\text{sparsing}\\ \text{topic}}}\Big)_+. \tag{3}$$

Коэффициенты β_w и α_t примем равными 1, $\forall w, t$. Коэффициенты регуляризации α, β и γ возьмём постоянными на протяжении всего прохода по коллекции. Их значения: $\alpha = 0.15, \beta = 0.009, \gamma = 7.8 \times 10^5$.

Результаты В таблице 1 приведены финальные значения функционалов качества после одного прохода по коллекции для моделей LDA и ARTM. Видно, что комбинация регуляризаторов разреживания и декорреляции улучшает качество результирующей модели с небольшими потерями перплексии.

Более подробно процесс обучения представлен на 1. На верхнем графике показано убывание перплексии и замеры разреженностей матриц Φ и Θ. На нижнем — усреднённые характеристики ядер тем. Видно, что LDA совершенно не способствует разреживанию и даёт менее чистые и контрастные ядра тем, чем ARTM.